BioSig: An Imaging Bioinformatic System for Studying Phenomics

B. Parvin, Q. Yang, G. Fontenay, and M.H. Barcellos-Hoff
Lawrence Berkeley National Laboratory
Berkeley, CA 94720
Web: http://vision.lbl.gov/Projects/BioSig

April 25, 2002

Abstract

Organisms express their genomes in a cell-specific manner, resulting in a variety of cellular phenotypes or phenomes. Mapping cell phenomes under a variety of experimental conditions is necessary in order to understand the responses of organisms to stimuli. Representing such data requires an integrated view of experimental and informatic protocols. BioSig provides the foundation for cataloging cellular responses as a function of specific conditioning, treatment, staining, etc. for either *in vivo* or *in vitro* studies. A data model has been developed to capture a wide variety of experimental conditions and map them to image collections and their computed high-level representations. Samples are imaged with light microscopy and each image is represented with an attributed graph. The graph representation contains information about cellular morphology, protein localization, and organization of the cells in the corresponding tissue or cultured colony. The informatics architecture is distributed and enables database content to be shared among multiple researchers.

1 Introduction

The challenge of the post-genomic era is functional genomics, i.e., understanding how the genome is expressed to produce myriad cell phenotypes. To use genomic information to understand the biology of complex organisms, one must understand the dynamics of phenotype generation and maintenance. A phenotype is the result of selective expression of the genome. It is an expression of the history of the cell and its response to the extracellular environment. In order to define cell "phenomes," it is necessary to track the kinetics and quantities of multiple constituent proteins, their cellular context and morphological features in large populations. Such studies should also include responses to stimuli so that functional models can be generated and tested. This paper focuses on an imaging bioinformatic system used for mapping cell phenomics.

Signaling between cells and their extracellular microenvironment has a profound impact on cell phenotype [5]. These interactions are the fundamental prerequisites for control of cell cycle, DNA replication, transcription, metabolism, and signal transduction. The ultimate decision of a cell to proliferate, differentiate or die is the response to integrated signals from the extracellular matrix, cell membrane, growth factors and hormones. Our current aim is to understand how ionizing radiation alters tissue homeostasis. This is achieved by studying the effect of low-dose radiation on the cellular microenvironment, inter-cell communication, and the underlying mechanisms. In turn, this information can then be used to more accurately predict more complex multicellular biological responses following exposure to ionizing radiation [1].

[*] Research was funded by the Low Dose Radiation Program of the Life Sciences Division, Medical Sciences Division, Mathematical, Information, and Computing Sciences Division, and the Director of Lawrence Berkeley Laboratory, of the U. S. Department of Energy under Contract No. DE-AC03-76SF00098 with the University of California. The publication number is LBNL-47456. Visit http://vision.lbl.gov or vision.lbl.gov/Projects/BioSig for more details.

For example, recent studies have shown that certain intracellular signaling pathways are linked via the cell adhesion system [6]. Cell adhesion is how a cell attaches itself via integral membrane receptors to the extracellular matrix. Experimentally manipulating extracellular matrix receptors affects cell shape, alters the response of cells to new stimuli, and modifies multicellular organization as a function of time [2]. Detailed analysis of these multidimensional responses (e.g., time and space) can be achieved using digital microscopy but is hampered by labor-intensive methods, a lack of quantitative tools, and the inability to index and access information. To motivate the practical aspects of the informatic system, consider the following. A typical study includes a number of genetically similar mice at different stages of their development: virgin, pregnant, lactate, and involution. In each category, mice are partitioned for treatment types (e.g., implant, radiation) that they will receive. Within each treatment population, mice are sacrificed at 1 hour, 4 hours, and 8 hours post treatment time. Tissues are then collected, sectioned, and coverslips are prepared for subsequent staining and imaging. The same experiment is then repeated for genetically altered mice for comparative analysis. It is clear that even such a simple study can generate a large number of images and annotation data to address cause and effect in the context of biological heterogeneity. The novelty of our system is a data model for capturing experimental annotations and variations, computational techniques for summarizing large number of images, and the distributed aspect of the architecture for distant collaboration.

The organization of this paper is as follows. Section 2 outlines various components of the informatic system. Section 3 summarizes the computational routines. Section 4 outlines two examples of phenotypic studies. Section 5 concludes the paper.

2 Informatics

Phenotyping has many degrees of freedom that should relate a particular quantitative result with (1) where a sample was obtained, (2) how it was conditioned, (3) how it was treated, etc. The BioSig informatic framework maintains these relations so that different experimental results can be compared for validation, exploratory analysis, and hypothesis testing. These relations encode a mapping between quantitative results to images and experimental annotations. The BioSig informatic system consists of three components. These include (1) data model, (2) presentation manager, and (3) query manager. These subsystems are decoupled for ease of development, testing, and maintenance. The data model captures experimental variations and their relationships, and maps sample preparation to images and their corresponding quantitative representation. The model is object oriented and allows bidirectional tracking between experimental annotation and computed representation. The presentation manager provides two distinct features: (1) mapping between the data model and the user interface, and (2) display functionality in terms of text, plots, and images. These features avoid hardwiring the user interface in favor of a more flexible model at run time. The query manager maps high-level user queries to the Java objects that implement the data model. The intent is to simplify and hide detailed manipulation of the database from the end users. The architecture for the informatics system is shown in Figure 1. It consists of a Web server (Java application servers) for direct communication to the database, computational services for image analysis, a CORBA bus, and a file system for storage of raw data. A key design decision has been not to provide a direct CORBA interface to the database at this point since current CORBA interfaces to the OO databases are weak and not well supported by various vendors. The database supports some computational functionalities on metadata; however, all image analysis operations are performed as a part of computational services.

2.1 Data model

The BioSig data model, shown in Figure 2, is object-oriented and provides navigational links between experimental variations, images, and quantitative analysis. In the actual implementation, each link often has a cardinality of more than one, and provides bidirectional tracking of information from any end point. The data model captures laboratory notebook information (e.g., information about various antibodies or treatments), experimental variation (e.g., type of external exposure and

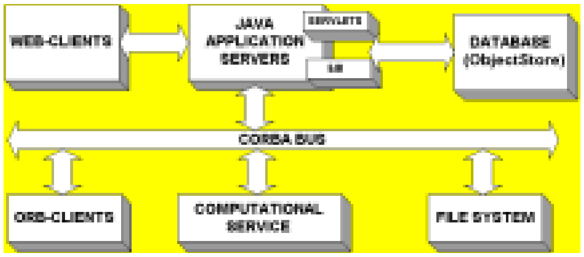


Figure 1: Distributed architecture for the informatics framework supports Web based access for distant collaboration.

its duration), images, computed morphological features, and protein co-localization features measured in each subcellular region. The data model supports both *in vivo* and *in vitro* studies, and it has been developed through repeated interviews with experimentalists who research different aspects of phenotyping that involves both fixed and live cell experiments. An *in vivo* study often consists of many animals for sensitivity and variational analysis. These animals are treated with a specific protocol, e.g., radiation or implants. Tissue sections are then prepared from an organ at a specific thickness, then stained with primary and secondary antibodies (a tool for studying protein) at specific dilutions. These stained samples are then imaged at different wavelengths (360 nm to 650 nm) and pertinent features are computed.

Another feature of BioSig is that it provides significant flexibility in the data model by augmenting each data object with a property object, which consists of name-value pairs that can be designated dynamically with general collection classes. The model is represented with XML, and software has been developed to convert the XML representation into the Java code that is required by the object oriented database. These capabilities allow evolution of the data model and accommodation of new features without requiring changes in the database access layers. In support of this evolutionary model, an interface has been developed to allow researchers to add new properties, specify their value types, and choose either to add them to instances on a predicate basis or to apply them globally.

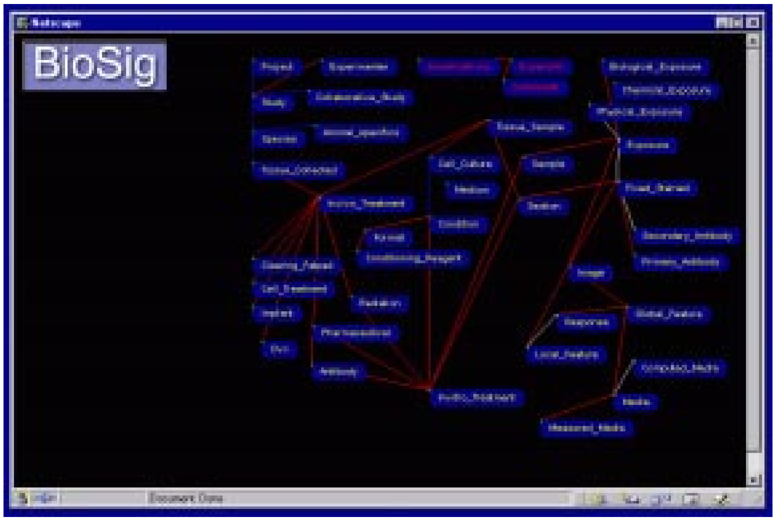


Figure 2: Coarse representation of the data model shown as a graph. The user can click on each object and view its content in more details.

2.2 Presentation manager

The BioSig presentation manager supports two features: (1) browsing the database and (2) visualizing the result of a query function. Browsing the database is performed against a predefined schema that captures annotation data, images, and corresponding high-level features. The data model, shown in Figure 2, is represented in XSD (XML schema), and the presentation manager constructs a view into the database using this representation and the corresponding style sheets (XSL) for browsing and updating. In this context, hardwiring of a GUI is bypassed in favor of a more flexible and dynamically generated user interface. In general, such a mapping may create a complex implementation issue. However, we have simplified the presentation system to allow browsing and updating one layer at a time. A layer refers to navigation between an object and other objects that are linked through association, aggregation, and inheritance. The presentation manager can display the result of a query function in either text or graphics. The graphics include dose-response plots and scatter diagrams of computed features as a function of independent variables. Examples of the presentation manager are shown in Figures 3 and 4.

2.3 Query manager

The BioSig query manager provides a set of predefined operators to assist in information visualization and hypothesis testing. These operators help to draw contrast between computed features and their corresponding annotation data and compute a variety of statistical measures such as analysis of variance and principal component analysis. In a sense, any query is performed by examples, and these operators translate a query into a Java program that manipulates the database to retrieve the required information. The object oriented database simplifies access of information such as analysis of variance since each computed feature has to be mapped to its source; e.g., animal or cell culture. An example of such a high-level operator includes correlation of a particular computed feature with respect to an independent variable; e.g., *correlate "organization" of an acinus between samples that have been treated with 2-Gy-levels of radiation and those that have not been radiated at all.* In this case, organization is a feature that quantifies global layout of a number of epithelial cells for a cell culture colony. Our system provides a visual query interface for mapping user queries to a set of database operations, computing the results, and transmitting them to the display manager. The actual computation may include analysis of variance or principal component analysis.

One particular type of operator is the "average" operator. In general, an experiment may include up to several hundred images that correspond to a particular tissue or cell culture that is fixed at a specific point in time. It is often necessary to visualize and represent a collection of images, at each time point, with two to three images that characterize the average behavior of that set. This average behavior is located through indexing of computed features, e.g., morphological or protein localization attributes.

3 Extraction of nuclei

Automatic delineation of nuclei from a large image collection is an important step in mapping protein localization into structural components for phenotypic studies. This process is also known as segmentation. In addition, for *in vivo* studies, one is often interested in protein expression for cells at a specific location in the tissue. Segmentation is a difficult problem due to noise, technical variation in sample preparation, and the fact that adjacent nuclei can overlap each other and form a clump. Noise can be random or speckled. While random noise is often due to CCD noise, speckled noise corresponds to tiny localized substructures (e.g., chromatin), which stand strongly against the diffused nuclear signature in the presence of a fluorescent dye. The segmentation technique is model-based, it incorporates a model for speckled noise, and it assumes that each 2D slice of a nucleus is locally quadratic along its boundary. The speckled noise can be extracted with elliptic features

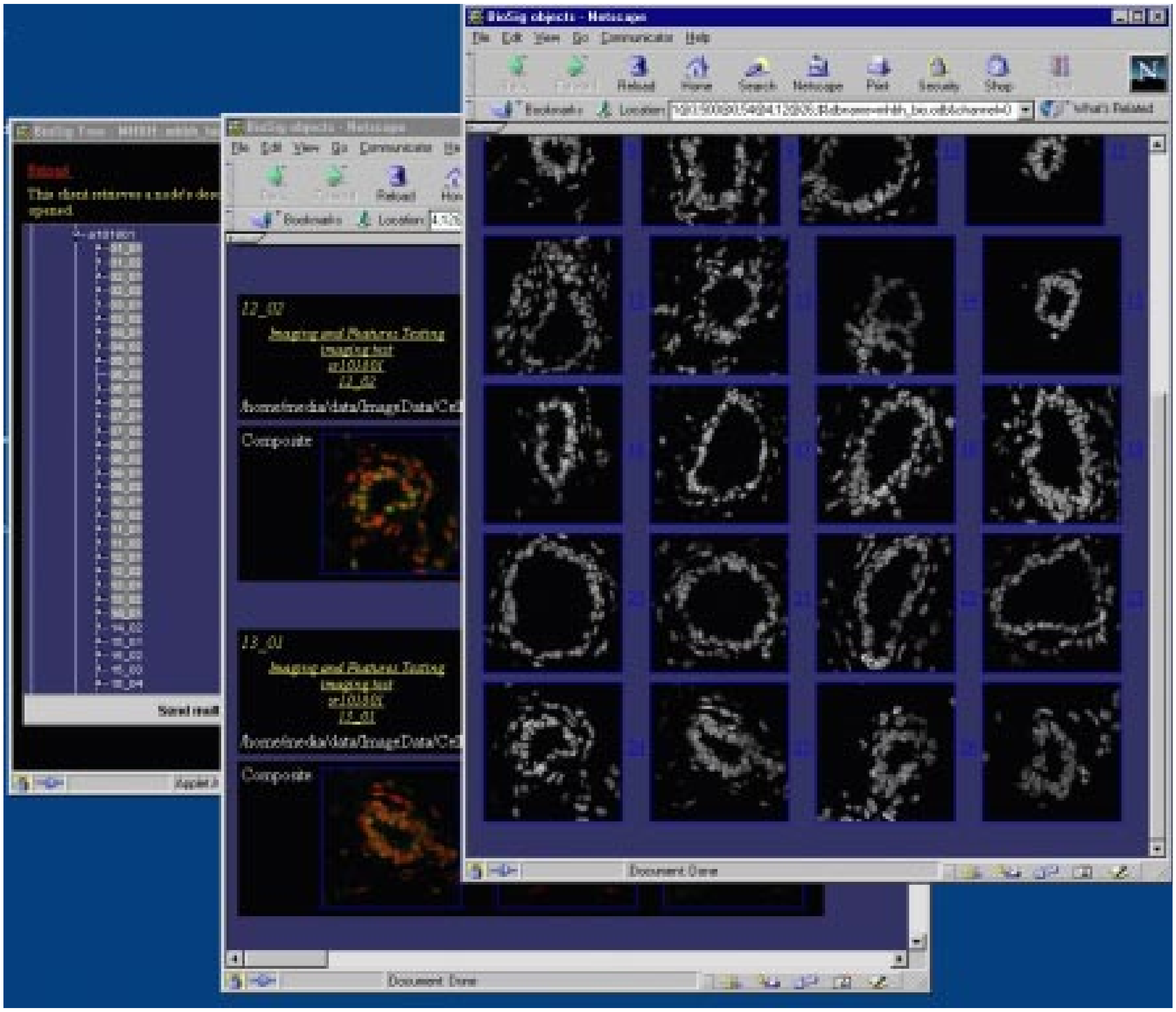
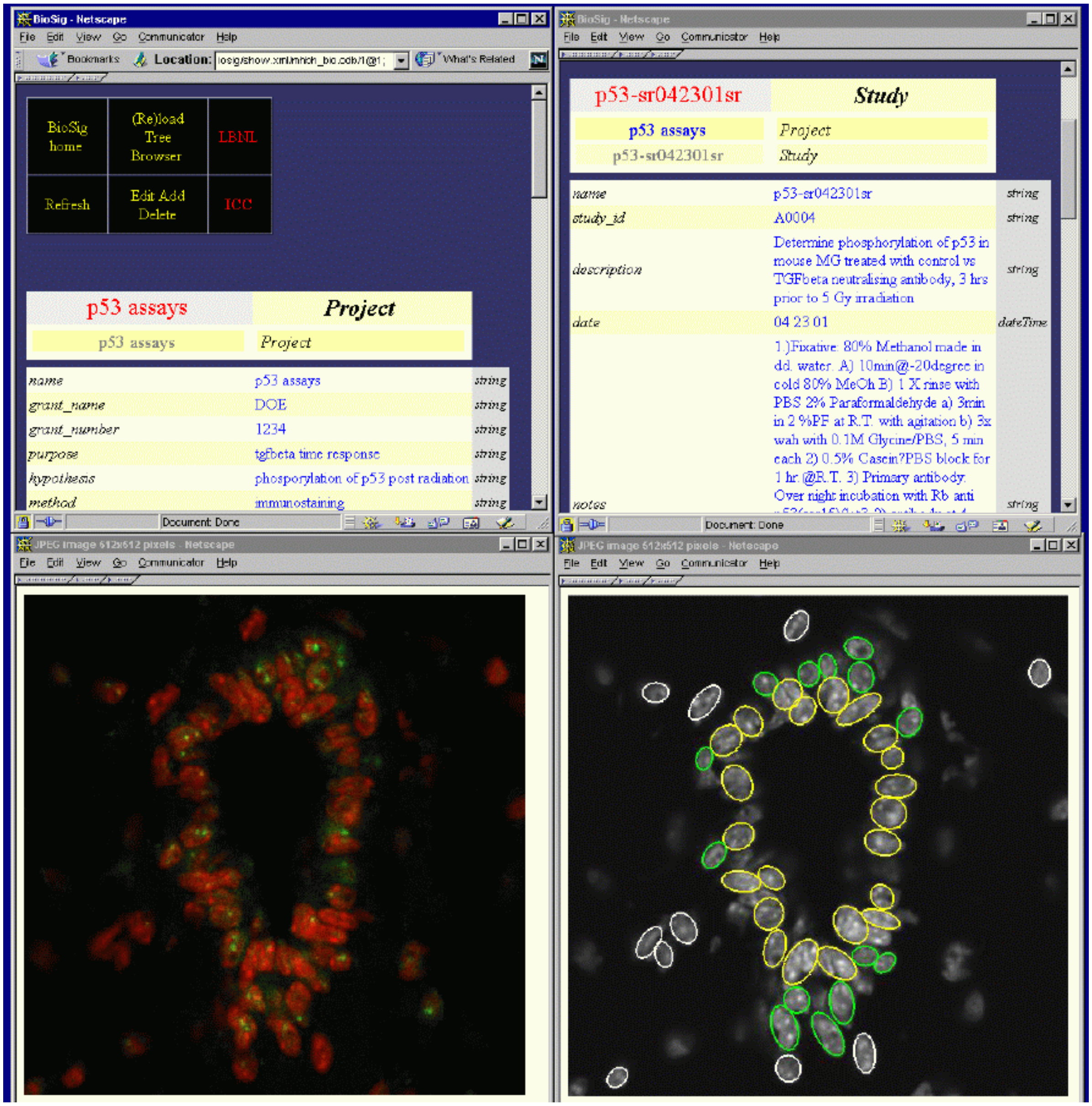


Figure 3: An example of the client's view of images and their annotation for *in vivo* experiments.



p53 assays	Project
p53 assays	Project

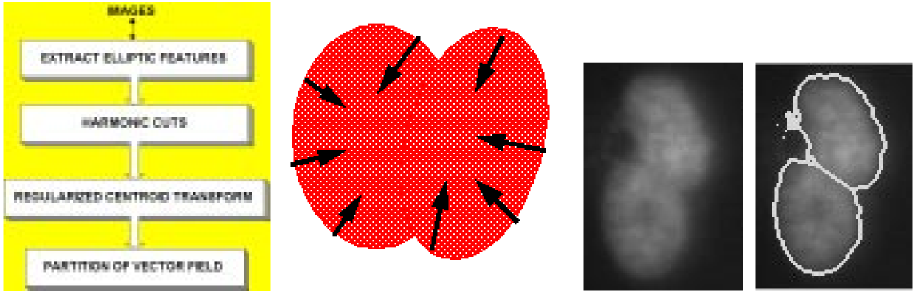
name	p53 assays	string
grant_name	DOE	string
grant_number	1234	string
purpose	tgfbeta time response	string
hypothesis	phosporylation of p53 post radiation	string
method	immunostaining	string

p53-sr042301sr	Study
p53 assays	Project
p53-sr042301sr	Study

name	p53-sr042301sr	string
study_id	A0004	string
description	Determine phosphorylation of p53 in mouse MG treated with control vs TGFbeta neutralising antibody, 3 hrs prior to 5 Gy irradiation	string
date	04 23 01	dateTime
notes	1.)Fixative: 80% Methanol made in dd. water. A) 10min@-20degree in cold 80% MeOh B) 1 X rinse with PBS 2% Paraformaldehyde a) 3min in 2 %PF at R.T. with agitation b) 3x wah with 0.1M Glycine/PBS, 5 min each 2) 0.5% Casein?PBS block for 1 hr. @R.T. 3) Primary antibody: Over night incubation with Rb anti	string

Figure 4: Client's view of the raw and processed images along with their annotation for an *in vivo* study. The image on the left combines two images that have been captured at two distinct frequencies. It is clear that protein expression (green color) is heterogeneous for cells in the immediate vicinity of each lumen. The image on the right shows segmentation and classification of each nuclei around the lumen.

and then interpolated with harmonic cuts (a noise removal step). Once noise is removed, touching nuclei are separated with a centroid transform as shown in Figure 5. Essentially, centroid transform collapses the content of each nucleus into its localized center of mass. The computational details can be found in our recent paper [8] or at our Web site.


IMAGES
EXTRACT ELLIPTIC FEATURES
HARMONIC CUTS
REGULARIZED CENTROID TRANSFORM
PARTITION OF VECTOR FIELD

(a) (b) (c) (d)

Figure 5: The segmentation protocol involves detection and removal of noisy regions with elliptic features, interpolating noisy regions with harmonic cuts, and separating touching compartments with centroid transform: (a) protocol for extracting delineating touching nuclei; and (b) graphical representation for evolution of centroid transform between two adjacent nuclei; (c) a pair of touching nuclei; and (d) result of partitioning adjacent nuclei.

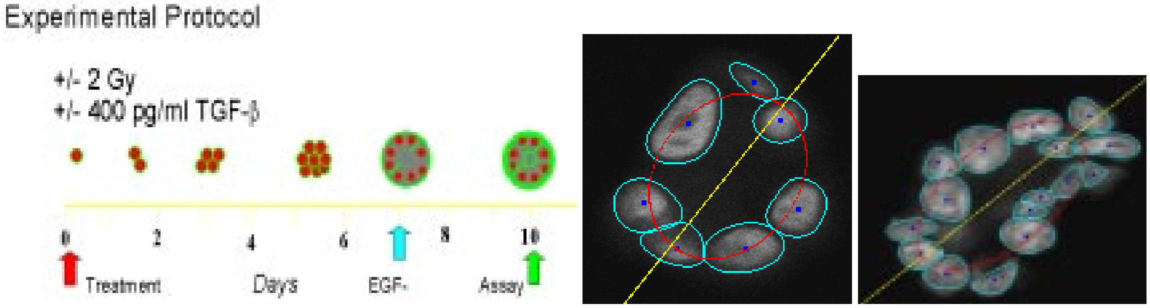
Phenotyping often involves multispectral imaging to couple morphological features with protein localization or physiological responses. A sample is usually tagged with a fluorescent dye and imaged at 360 nm to reveal nuclear formation (shape and organization). Protein expression is imaged at other excitation wavelength; e.g., 490 nm and 570 nm. Nuclei are delineated using the technique described in section 3 and then represented with an ellipse as well as hyperquadrics. Hyperquadrics is a parametric representation of an arbitrary shape with a series polynomials, and the details can be found in our earlier paper [4]. The underlying structure is then used as a mask to characterize protein expression at a specific excitation frequency. Finally, the collection of cells is represented as a region adjacency graph, where each node corresponds to a cell and each edge corresponds to the relationship between neighboring cells. For *in vivo* studies, each nucleus in the image is further classified with respect to the position in the lumen. This classification is based on locating the lumen inferred from the region adjacency graph and labeling the cells in the immediate vicinity of the lumen as the lumenal epithelial cells. Figure 4 shows examples of ellipse fitting and nuclear classification with coded colors. Note that protein expression (represented as green color on the left sub-image) is neither diffused in the nuclear region nor homogeneous in cells with the same classification. Although segmentation and classification is automatic, tools are provided to edit the classification results.

4 Applications

Two applications are included here to show the use cases of BioSig. The first one corresponds to cell culture studies involving cell-cell communication and adhesion. The second one provides the basis for establishing a link between extracellular events and intra-cellular signaling for normal (wild type) versus genetically altered animals.

4.1 *In vitro* studies

During cell culture studies, a single lumenal epithelial cell divides to form a hollow sphere known as an acinus. This process often takes 10 days, when at different time points, the microenvironment is disrupted to study cell-to-cell communication. To determine whether low-dose radiation promotes aberrant extracellular matrix (ECM) interactions, we have utilized BioSig to examine integrin and E-cadherin localization in preneoplastic human cells surviving radiation. Integrins are a family of epithelial receptors for the ECM, while E-cadherin maintains normal cell-cell interactions and architecture. We used the HMT-3522 (S1) human breast cell line cultured within a reconstituted ECM. These cells are genomicaly unstable but phenotypically normal in that they recapitulate normal mammary architecture in the form of a multicellular, three-dimensional acinus [7]. These clusters express integrins in a polarized fashion and develop an organized ECM over the course of 7-10 days in culture. The intent is to examine the consequences of exposing these cells to ionizing radiation and a protein modifier known as TGFβ as shown in Figure 6a. Antibodies to E-cadherin, beta 1 integrin or alpha 6 integrin were detected using a green fluorescent label, while nuclei were counter-stained with a red fluorescent DNA dye. Cells that survived either 2 Gy or TGFβ (400 pg/ml) showed decreased beta 1 or alpha 6 integrin localization, respectively. However, when cells were exposed to both radiation and TGFβ, additional perturbations were noted. The clusters were disorganized (see Figure 6c), did not polarize the integrins at the cell surface, and failed to express E-cadherin, indicative of a lack of structural organization.



(a) (b) (c)

Figure 6: Organization of a colony as a result of low-dose radiation and TGFβ treatment indicates lack of symmetry around the lumen. Nuclei are segmented, represented with hyperquadrics, and symmetry is measured by fitting an ellipse to all nuclei: (a) experimental protocol; (b) an untreated sample maintains symmetry along the lumen; and (c) a treated sample loses its symmetric organization.

4.2 *In vivo* studies

One of the most rapid cellular responses to low-dose radiation is the activation of the transcription factor p53 (a DNA repair molecule), whose abundance and action dictates in individual cellular consequences regarding proliferation, differentiation, and apoptosis. Described as the guardian of the genome by *Science* in 1995, p53 is one of the most rapid cellular responses to radiation. Activation of p53 allows it to bind to DNA and to transactivate target genes. A major cellular function of the p53 tumor suppressor protein is its role in promoting genome integrity. Whereas *intracellular* radiation-induced mediators of p53 stability have been the subject of intense study, little is known about the *extracellular* factors that affect the p53 response to ionizing radiation. A number of striking similarities exist between p53 and TGFβ: both regulate complex cellular decisions regarding cell fate [3], both are induced by a variety of damage and specifically ionizing radiation, and both are rapidly activated and exist in latent forms. In the present study, we used p53 antibodies that bind to a phosphorylated form of the protein that is induced upon radiation exposure. The significance of this study is that TGFβ is extracellular while p53 is intracellular. Confocal microscopy is used to collect the distribution of p53 immunoreactivity. Nuclear features such as shape, size, volume, relative location, and intensity are computed and stored in the database. These features are then used to track the level and distribution of p53 within specific tissue compartments. The result is shown in Figure 7, where BioSig provides a visual representation of p53 expression in three categories of nuclei (red for lumenal epithelial, cyan for myo-epithelial, and blue for stromal cells) for a population of images from wild-type tissue sections.

Next, an experiment was designed to study the impact of TGFβ on the p53 as a result of an external exposure and different strands of mice, which may genetically altered. Normal mice (control animals) were exposed to low-dose radiation, tissues were collected, samples were treated with antibodies, and a large number of images were produced. Genetically altered mice, with only one copy of TGFβ (as opposed to two), were also externally exposed, etc. The protocol was repeated without any external exposure on both strands of mice. The experiment produced a large amount of data that is archived in the database along with their annotations. The results indicate that p53 is expressed less in genetically altered mice, thus, a link between extracellular condition and intracellular event is made.

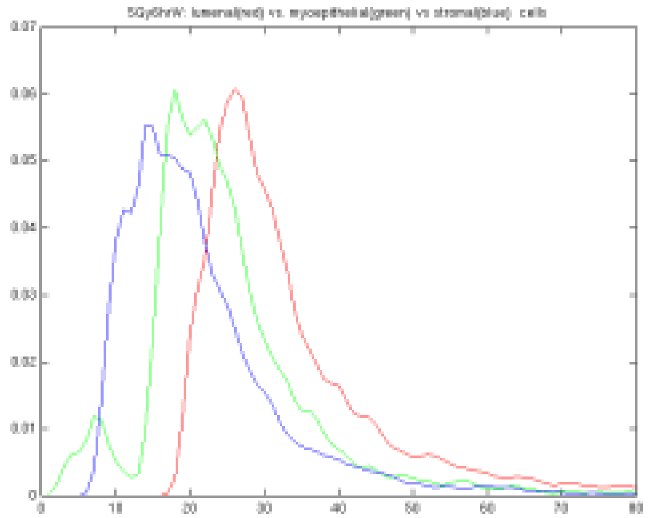


Figure 7: Population studies for p53: probability density functions for response of p53 in each cell type (red: lumenal-epithelial, green: myo-epithelial, blue: stromal).

5 Conclusion

In the post-genome-sequencing era, quantitative imaging of complex biological materials is a critical problem. Currently, sequential measurements obtained with different microscopy techniques preclude detailed analysis of multidimensional responses. Quantification of spatial and temporal concurrent behavior of multiple markers in large populations of multicellular aggregates is hampered by labor-intensive methods, a lack of quantitative tools, and the inability to index information.

There are several thousand antibodies and reagents for differentiating specific protein components of cells. Some antibodies can additionally discriminate between functional variants of a protein caused by modifications such as phosphorylation status, protein conformation and complex formation. Of the intracellular proteins, a large number are involved in signaling pathways. These pathways are currently not well understood due to the complexity of the potential events, the potential for multiple modifications affecting protein function, and lack of information regarding where and when a protein is actively participating in signaling. Inherent biological variability and genomic instability are additional factors that support the requirement for large population analysis. The BioSig informatics approach to microscopy and quantitative image analysis is being used to build a more detailed picture of the signaling that occurs between cells as a result of an exogenous stimulus such as radiation or as a consequence of endogenous programs leading to biological functions. The details of data model, usage of the database, and methods for importing legacy data into the database are posted on the Web. Public access to certain aspect of the database will soon be facilitated.

References

[1] M.H. Barcellos-Hoff. How do tissues respond to damage at the cellular level? the role of cytokines in irradiated tissues. *Radiation Research*, 150:109–120, 1998.

[2] F.G. Giancotti and E. Ruoslahti. Integrin signaling. *Science*, 285:1028–1032, 1999.

[3] A.J. Levine. P53, the cellular gatekeeper for growth and division. *Cell*, 88:323–331, 1997.

[4] B. Parvin, G. Cong, J. Fonteny, J. Taylor, R.L. Henshall, and M.H. Barcellos-Hoff. Biosig: a bioinformatic system for studying the mechanism of inter-cell signaling. In *IEEE International Symposium on Bio-Informatics and Biomedical Engineering*, pages 281–288, 2000.

[5] C.D. Roskelley, A. Srebrow, and M.J. Bissell. A hierarchy of ecm-mediated signalling regulates tissue-specific gene expression. *Current Opinion in Cell Biology*, 7(5):736–747, 1995.

[6] F. Wang, V.M. Weaver, O.W. Petersen, C.A. Larabell, S. Dedhar, P. Briand, R. Lupu, and M.J. Bissel. Reciprocal interactions between beta 1-integrin and epidermal growth factor receptor in three-dimensional basement membrane breast cultures: A different perspective in epithelial biology. *Proceedings of the National Academy of Sciences of United States of America*, 95(25):14821–14826, 1998.

[7] V.M. Weaver, A.H. Fischer, O.W. Petersen, and M.J. Bissel. The importance of the microenvironment in breast cancer progression: recapitulation of mammary tumorigenesis using a unique human mammary epithelial cell model and a three-dimensional culture assay. *Biochemical Cell Biology*, 74(12):833–51, 1996.

[8] Q. Yang and B. Parvin. Harmonic cuts and regularized centroid transform for localization of subcellular structures. In *Proceedings of the International Conference on Pattern Recognition*, 2002.